Trend Analysis of E-Commerce Data using Hadoop Ecosystem

Rama Satish K. V.
Research Scholar of VTU
RNSIT Research Center,
Bangalore

N. P. Kavya
Professor and Manager - HR
RNSIT, Bangalore,
India

ABSTRACT
Trend Analysis is the custom of collecting information and attempting to spot a trend, or pattern, in the information. Trend analysis is often used to estimate future events, it could be used to approximate uncertain events in the past. Technical analysts and Technicians also uses market indicators of many types. Processing or analyzing the Trend in huge amount of data or extracting meaningful information is a challenging task. As the enterprises faced issues of gathering large chunks of data and analyzing the Trend .They found that the, data cannot be processed using any of the existing centralized architectures. One of the best open source tools used in the market to harness the distributed architecture in order to solve the data processing and analyzing problems is Apache Hadoop and Hive for querying best results. This paper addresses an experimental work on Trend analysis problem of big data and its optimal solution using Hadoop ecosystem, using parallel processing framework to process large data sets using Map Reduce programming and Apache Hive is a data warehouse infrastructure which is built on top of Hadoop for providing data summarization, querying and analysis.

Keywords
Trend Analysis, E-Commerce data, Apache Hadoop, Hive.

1. INTRODUCTION
Trend Analysis is the practice of collecting information and to spot a trend or pattern, in the information. Trend analysis is mostly used to predict upcoming events, it could be used to analyze uncertain events in the past, like how many ancient kingdoms ruled between dates, based on data such as the average years which other known kings ruled. Trends analysis could be a motivation, but could also determine the risks involved while introducing new products. Research in trend analysis is very complex and extremely difficult to identify and analyze the future trends. Trends are changes in societies that occur over longer periods of time. Trends are not only shifts in people's preferences like fashion, dance, drama or music, but are also shifts in larger areas such as the economy, politics, and technology. Trends are very important source of inspiration for thinking up new products. Trends analysis is often used as part of a strategic planning process. Trends are used to identify customer needs, which a company can meet with their new products or services. Trends analysis is preceded by trends watching, by which we mean the identifying, gathering and reporting of trend information without giving insight into the possible consequences.

Apache Hadoop is a well-known project that includes open source implementations of a distributed file system and a MapReduce parallel processing framework that were inspired by Google's GFS and MapReduce projects. The emergence of the open source Hadoop system eliminates the technical barrier to cloud computing. Several rising stars of international IT companies, like Whatsapp, Facebook and Twitter, are dedicated to making contributions to the Hadoop community as well as deploying and using this system to building their own cloud computing systems. After several years of development, Hadoop gradually forms a cloud computing ecosystem consisting of a set of technical solutions which include the HBase distributed database, Hive distributed data warehouse, ZooKeeper coordination services for distributed applications and etc. All the components are built on the top of low-cost commercial hardware, with the extensive availability and fault-tolerance, which makes Hadoop gradually become the mainstream of a commercial implementation as a cloud computing technology. One of the significant designed features of the Hadoop system is high throughput which is extremely suitable for handling large-scale data analysis and processing problems.

In this paper we analyze the trend of E-Commerce web pages traffic logs of www.amazon.com. The traffic department generates a daily log of programming element and enters them into a computer system that will help to generate the daily logs.

2. RELATED WORK
Trend Analysis have been made in many areas. Trend analysis of keyword frequency on online Novels as in [5], Trends in Temperature and extreme temperature in particular region in [6], Trend analysis of aircraft refrigerator based on rough set algorithm[7],[8] shows how it is used in software testing, Trend analysis of personal health record [9] and in field of business and many more. Trend Analysis is the practice of collecting information and trying to spot a pattern or trend, in the information. Trend analysis is mainly used to predict upcoming events, it could be used to estimate uncertain events in the past, such as how many ancient kings probably ruled between two dates, based on data such as the average years which other known kingdoms ruled. In some fields of study, the term "trend analysis" has more formally defined meanings, in case of project management trend analysis is a mathematical technique that uses historical results to predict future outcome. This is achieved by tracking variances in cost and schedule performance. In this, it is a project management quality control tool. In statistics, trend analysis often refers to techniques for extracting an underlying pattern of behavior in a time series which would otherwise be partly or nearly completely hidden by noise. A simple description of these techniques is trend estimation, which can be undertaken within a formal regression analysis. Trend analysis also refers to the science of studying changes in social patterns, including fashion, technology and consumer behavior. In this paper we are doing Trend analysis of E-Commerce web pages. Trending Topics would continue to show a range of stories when this type of event happens. Some of the automated news sites like Techmeme or Google News display clusters of

rising articles across a number of topics to provide a better mix of content on the homepage. For example, Techmeme shows a ranked list of related blog posts that link back to the lead article for each story. We can build a simple version of this functionality with Hadoop by combining the article trend estimates we computed in the previous post with the E-Commerce web pages link graph. E-Commerce web portals provides a periodic database dump which includes a file with the outgoing page links for each article.

The current trend calculations are run with Hadoop Streaming and Hive. The output obtained by these Hadoop jobs is loaded into MySQL and indexed to power the live site. A wide variety of techniques and technologies has been developed and adapted to aggregate, manipulate, visualize, and analyze big data. These techniques and technologies draw from several fields including computer science, applied mathematics, statistics and economics. This means that an organization that intends to derive value from big data has to adopt a flexible, multidisciplinary approach. Dealing with big data requires two things Inexpensive, reliable storage and new tools for analyzing unstructured and structured data. Apache Hadoop is a powerful open source software platform that addresses both of these problems.

2.1 Apache Hadoop

Hadoop is a software framework that supports data-intensive distributed applications in [1],[2],[3] . It enables applications to work with thousands of computational independent computers and peta bytes of data. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Hadoop is completely written in Java and is cross platform. Hadoop enables the development of reliable, scalable, efficient, economical and distributed computing using very simple Java interfaces - massive parallel code without the pain!

2.2 Hadoop Inspired by Google File System

Hadoop was originally built as an infrastructure for the Nutch project, which crawls the web and builds a search engine index for the crawled pages. Nutch was started in 2002, and a working crawler and search system quickly emerged. However, they realized that their architecture wouldn't scale to the billions of pages on the Web. Help was at hand with the publication of a paper in 2003 that described the architecture of Google's distributed file system, called GFS, which was being used in production at Google. During 2004, they set to write an open source implementation, the Nutch Distributed File System (NDFS). In 2004, Google published the paper that introduced MapReduce to the world. In 2005, the Nutch developers had a working MapReduce implementation in Nutch major Nutch algorithms had been ported to run using MapReduce and NDFS.

2.3 Hadoop Distributed File System

Hadoop includes a fault tolerant storage system called the Hadoop Distributed File System, or HDFS. HDFS is to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop is ideal for storing large amounts of data, like terabytes and petabytes, and uses HDFS as its storage system. HDFS connect nodes contained within clusters over which data files are distributed. Helps to access and store the data files as one seamless file system. HDFS has master/slave architecture. An HDFS cluster consists of a single Name node, a master that manages the file system namespace and regulates access to files by clients. In addition, there are a number of Data nodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. HDFS exposes a file system namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks and these blocks are stored in a set of Data nodes. The Name node executes some of file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to Data nodes. The Data nodes helps for serving read and write requests from the file system's clients. The Data nodes also perform block creation, deletion, and replication upon instruction from the Name node.

2.4 MapReduce - Programming Model

MapReduce is a linearly scalable programming model. The Figure 2.1 shows the basic diagram of map reduce working. The programmer writes two functions - a map function and a reduce function - each of which defines a mapping from one set of key-value pairs to another is shown in [14]. These functions are obvious to the size of the data or the cluster at which they operate, so they can also be used even for a small dataset too and for a massive one. More importantly, when the size of the input data is doubled, a job will run twice as slow. But the size of the cluster is increased, a job will run as fast as the original one. Hadoop's MapReduce and HDFS use simple robust framework runs on commodity hardware to deliver high data availability and to analyze enormous amounts of information quickly. Hadoop offers enterprises a powerful new tool for handling big data. Hadoop creates clusters of machines and coordinates work among them. Clusters can be built with inexpensive computers. If one fails, Hadoop continues to operate the cluster. The main limitation of the MapReduce paradigm is that each Map and Reduce task must not depend on any data generated in other Map or Reduce tasks of the current job, as user cannot control the order in which they execute. Although applicable to a wide variety of problems, there are problems to which the MapReduce is not directly applicable. These are recursive computations, and algorithms that depend on shared global state.

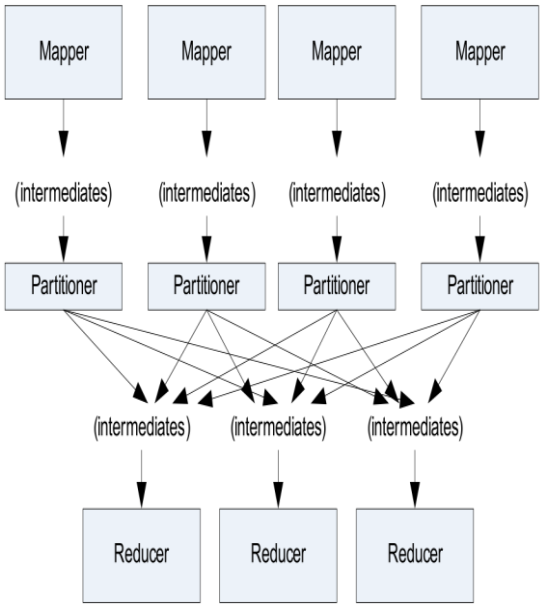

Figure 2.1: Map Reduce Framework

2.5 Apache Hive

Apache Hive is an open-source data warehouse system for querying and analyzing large datasets stored in Hadoop files in [4]. Hadoop is a framework for handling large datasets in a distributed computing environment. HiveQL is the Hive query language. Like all SQL dialects in widespread use, it doesn't fully conform to any particular revision of the ANSI SQL standard.

2.6 Sqoop

Sqoop is a tool designed to transfer data between Hadoop and relational databases. You can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the data back into an RDBMS. Sqoop in [10] automates most of this process, relying on the database to describe the schema for the data to be imported. Sqoop uses MapReduce to import and export the data, which provides parallel operation as well as fault tolerance. Allows data imports from external datastores and enterprise data warehouses into Hadoop Parallelizes data transfer for fast performance and optimal system utilization Copies data quickly from external systems to Hadoop.

3. SYSTEM ARCHITECTURE

E-Commerce web page name has been requested for the past two years, for each day, month and for the previous 60 or 90 days. These figures do not reflect the number of unique visitors a page has received.[1] They also do not reflect how often the page has been read or even viewed. The Figure 3.1 shows the overall system architecture of our work. The pageview stats tool is available from any page, in two ways: 1) see the toolbox in the sidebar, which shows page information; the external link is in the last section of page information; and 2) look behind the page's history tab and select page view statistics.

3.1 Preprocessing the Raw Data

The daily page view charts on the site were created by running an initial MapReduce job on 1TB of hourly traffic logs are collected.

The log files are named with the date and time of collection. Individual hourly files are around 90 MB when compressed, so one month of compressed data takes up about 64 GB of space. Each line in a pagecount file has four fields: projectcode, pagename, pageviews, and bytes:

$grep'^enBarack'pagecounts20090521100001

en Barack 8 1240112

en Barack%20Obama 1 1167

en Barack_H._Obama 1 142802

en Barack_H_Obama 3 428946

en Barack_H_Obama_Jr. 2 285780

en Barack_Hussein_Obama,_Junior 2285606

en Barack_O%27Bama 1 142796

en Barack_Obama 701 139248439

en Barack_Obama%27s_first_100_days 2 143181

en Barack_Obama,_Jr 2 285755

Many records in the log file are actually E-Commerce web pages redirects that point to other articles in the E-Commerce web pages "Pages" table. Hadoop processes and cleans the page names and perform a join in Hive against the contents of a MySQL redirect table to find the true E-Commerce web portal page_id for each page title. The first MapReduce pass restricts page views to a subset of English E-Commerce web pages, filters out bad records, and then sums hourly page views keyed by article-date. If the date is noticed it is not actually in the raw log data, but it is a part of the filename. We can access this parameter in this streaming script using a Hadoop. Second MapReduce pass maps the daily aggregations by article name

Barack_Obama 20090422 129

Barack_Obama 20090419 143

Barack_Obama 20090421 163

Barack_Obama 20090420 152

These records are merged at the reducers to generate a daily time series for each article in serialized JSON format.

Format: article\tdates\tpagecounts\ttotal_pageviews

'Barack_Obama\t[20090419,20090420,20090421,20090422]\t[143,152,163,129]\t587′

These records are joined with the E-Commerce web page ID using Hive, and the resulting output is loaded into MySQL where it is indexed for fast lookups by the web application.

3.2 Daily Trend Estimation

After the historical timeline aggregation was complete, running a daily batch job in Hive to aggregate recent log data and detect topics trending over the previous 24 hours. This MapReduce job only operates on the last 30 days of data for trend estimation, so it is less resource intensive than a pass over the full E-Commerce web log dataset.

3.3 Querying for results using Hive

Doing Joins in MapReduce can be a bit of a pain, Hive hides a lot of tedious details behind a simple SQL like syntax most developers are familiar with. Behind the scenes this compiles down to optimized MapReduce code executed by Hive. Import and export of data is also very natural and will be familar to MySQL users. Download Page table dump and the Redirect table dump separately, Import Page and Redirect table into mysql and Using Sqoop import both tables into Hive. Join Page table and Redirect table to get aggregate views of a True Page Title.

4. EXPERIMENTAL SETUP

The trending topics application identifies recent trends on the web by periodically launching Cloudera's Distribution for Hadoop to process E-Commerce web log files. The daily page view charts on the site were created by running an initial MapReduce job on GB's of hourly traffic logs collected from Wikipedia's mysql dumps. We run a job on the http://www.trendingtopics.org server to fetch the latest log files every hour and store a copy on mysql db for processing by Hadoop and we use Sqoop to load the data into hive.

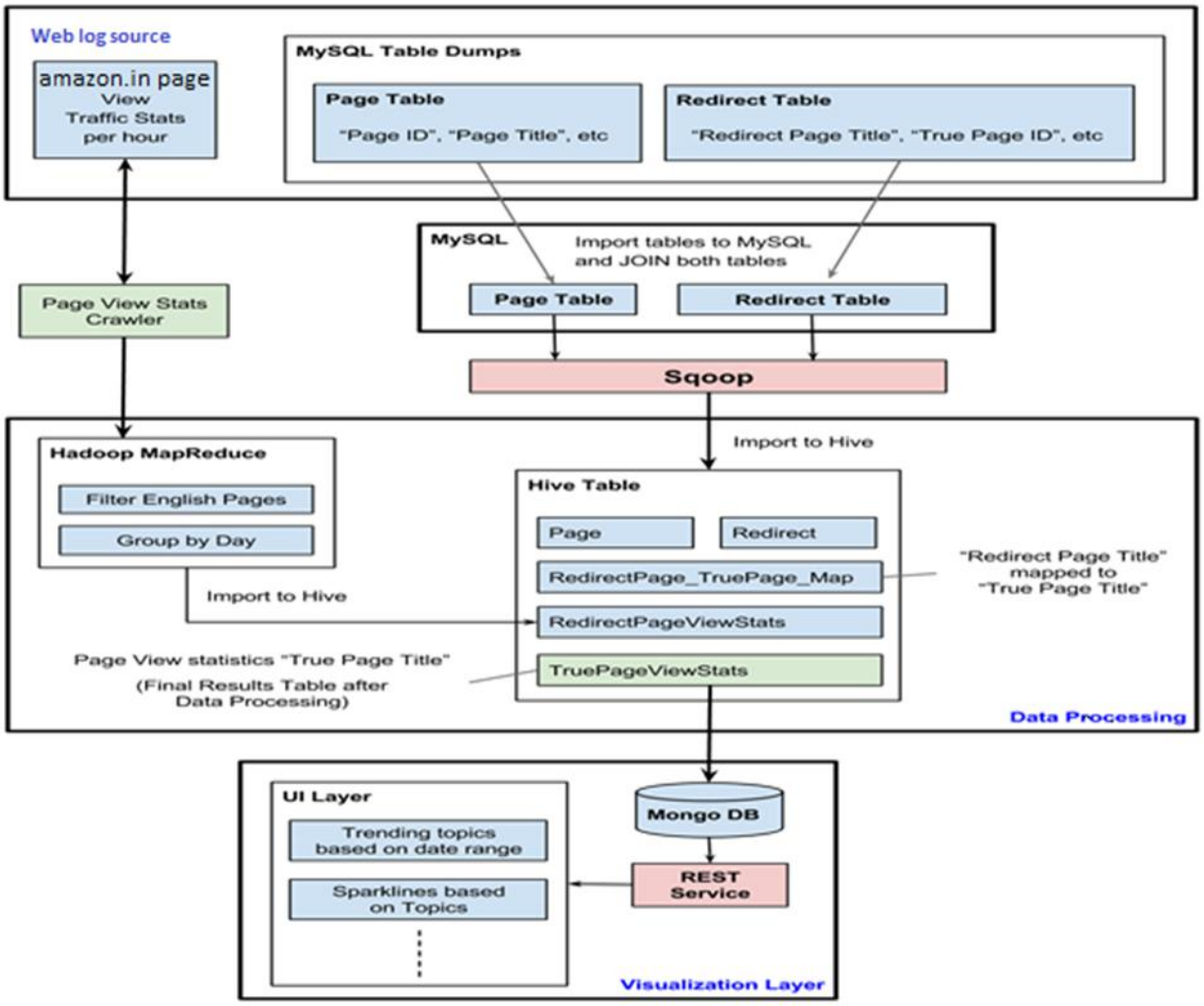

Fig 1: If necessary, the images can be extended both columns

5. RESULTS

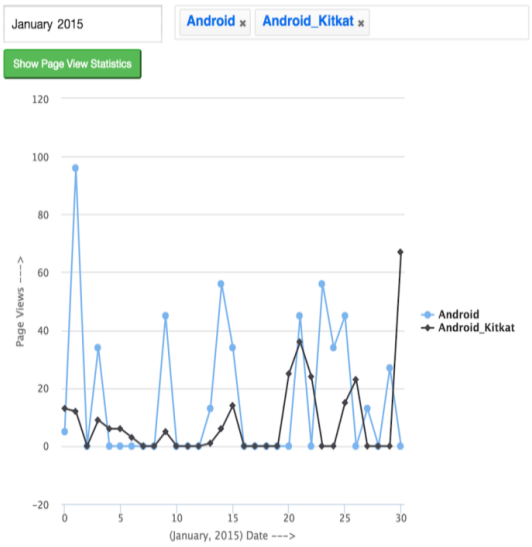

Figure 5.1 Output Result

The cluster runs a Hive batch job that analyzes hourly pageview statistics for millions of E-Commerce web pages, and then loads the resulting trend parameters into the application's NoSQL MongoDb after which output is populated and visualized on high charts to view the ranks of trending topics as shown in the Figure 4.1. For performing the big data experiments, setup of Hadoop data cluster and Hadoop Distributed File System (HDFS) for storage was used. Before working on multi-node cluster, single node cluster was first configured and tested. We configured our cluster to run map reduce jobs for finding trends in log data with Hive.Input and output data for the Map/Reduce programs is stored in HDFS, while input and output data for the data-parallel stack-based implementation is stored directly on the local disks. The software used to setup these hosts are Sun Java 1.7, Cloud era quick start vm 5.1.0 and for visualization layer we have used D3.js.

6. CONCLUSION

Trend Analysis is the practice of collecting information and attempting to spot a pattern, or trend, in the information. The aim of this project is to analyze trend of E-Commerce web pages given by the E-Commerce web traffic logs. The current trend calculations are run with Hadoop Streaming and Hive. The output produced by these Hadoop jobs is loaded into MySQL and indexed to power the live site. Doing Joins in MapReduce can be a bit of a pain,Sqoop is used for importing data to Hive. Apache Hive an open-source data warehouse system for querying and analyzing large datasets stored in Hadoop files.

7. REFERENCES

[1] Impetus white paper, March, 2011, "Planning Hadoop/NoSQL Projects for 2011" by Technologies

[2] ChenHauWang Dept. of Comput. Sci., Nat. Chiao-Tung Univ., Hsinchu, Taiwan Ching TsorngTsai ; Chia Chen Fan ; Shyan Ming Yuan A Hadoop Based Weblog Analysis System.

[3] Rama Satish K V ; N P Kavya "An approach to optimize QOS Scheduling of MapReduce in Big Data", International Journal of Engineering Research and Technology, Volume 2, Issue 11, May 2014.

[4] Fuad, A. ; Erwin, A. ; Ipung, H.P.Information, Communication Technology and System (ICTS), 2014 International Conference on DOI:10.1109/ICTS.2014 .7010600 Publication Year: 2014 , Page(s): 297 – 302.

[5] Rama Satish K V ; N P Kavya, "Big Data Processing with harnessing Hadoop - MapReduce for Optimizing Analytical Workloads", Proc. of IEEE International Conference, Mysore, INDIA, 27-29, November 2014.

[6] Dan Meng ; Huili Gong ; Xiaojuan Li ; Demin Zhou Geoinformatics (GEOINFORMATICS), 2012 20th International Conference on DOI:10.1109/Geoinformatics.2012.6270336 Publication Year: 2012 , Page(s): 1 – 6.

[7] Jianguo Cui ; Pengyuan Zhao ; Shiliang Dong ; Liqiu Liu ; Rui Lv ; Zhonghai Li Electrical and Control Engineering (ICECE), 2011 International Conference on DOI: 10.1109/ICECENG.2011.6057830Publication Year: 2011 , Page(s): 3339 – 3342.

[8] Meng Gui-fang ; Cheng Wan-li ; Zhu Wei Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on DOI: 10.1109/ICCSN .2011.6013692 Publication Year: 2011 ,Page(s): 191 - 194

[9] "Why Big Data is a must in E-Commerce", Guest post by Jerry Jao, CEO of Retention Science. http://www.bigdatalandscape.com/news/why-big-data-is-a-must-in-ecommerce.

[10] Apache sqoop official website: http://sqoop.apache.org/docs/

[11] Rama Satish K V and N P Kavya, "A New Efficient Cloud Model for Data Intensive Application", [GJCST] Global Journal of Computer Science and Technology: Distributed and Cloud Computing, March 2015.

[12] Yogesh Pingle, Vaibhav Kohli, Shruti Kamat, Nimesh Poladia, (2012) "Big Data Processing using Apache Hadoop in Cloud System", National Conference on Emerging Trends in Engineering & Technology.

[13] Tom White, (2012) "Hadoop: The Definitive Guide. O'Reilly", Scbastopol, California.

[14] Jeffrey Dean and Sanjay Ghemawat., (2004) "MapReduce: Simplified Data Processing on Large Clusters", Google Research Publication.